Reduced minimax state estimation

Vivien Mallet, Sergiy Zhuk

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Reduced minimax state estimation

Vivien Mallet — Sergiy Zhuk

N° 7500 — version 1.0

initial version December 2010 — revised version Décembre 2010

Observation and Modeling for Environmental Sciences

ISSN 0249-6399 ISRN INRIA/RR--7500--FR+ENG

INSTITUT NATIONAL
DE RECHERCHE
EN INFORMATIQUE
ET EN AUTOMATIQUE

INRIA

centre de recherche
PARIS - ROCQUENCOURT

Reduced minimax state estimation

Vivien Mallet [*] [†], Sergiy Zhuk [*] [‡]

Theme : Observation and Modeling for Environmental Sciences
Équipe-Projet Clime

Rapport de recherche n° 7500 — version 1.0 — initial version December 2010 — revised version Décembre 2010 — 23 pages

Abstract: A reduced minimax state estimation approach is proposed for high-dimensional models. It is based on the reduction of the ordinary differential equation with high state space dimension to the low-dimensional Differential-Algebraic Equation (DAE) and on the subsequent application of the minimax state estimation to the resulting DAE. The DAE is composed of a reduced state equation and of a linear algebraic constraint. The later allows to bound linear combinations of the reduced state's components in order to prevent possible instabilities, originating from the model reduction. The method is robust as it can handle model and observational errors in any shape, provided they are bounded. We derive a minimax algorithm adapted to computations in high-dimension. It allows to compute both the state estimate and the reachability set in the reduced space.

Key-words: minimax, reduction, differential-algebraic equations, estimation, filtering

* INRIA
† CEREA, joint laboratory École des Ponts ParisTech - EDF R&D, Université Paris-Est
‡ Taras Shevchenko National University of Kyiv

Centre de recherche INRIA Paris – Rocquencourt
Domaine de Voluceau, Rocquencourt, BP 105, 78153 Le Chesnay Cedex
Téléphone : +33 1 39 63 55 11 — Télécopie : +33 1 39 63 53 30

Filtrage minimax réduit

Résumé : Nous introduisons une méthode de filtrage dédiée aux modèles de grande dimension et fondée sur une approche minimax réduite. La méthode repose sur une reformulation du problème de grande dimension en une équation différentielle algébrique de petite dimension sur laquelle un filtre minimax est appliqué. L'équation différentielle algébrique se décompose en une équation sur un état réduit et une contrainte algébrique linéaire. Cette dernier permet de borner des combinaisons linéaires des composantes du vecteur d'état réduit, ce qui élimine des instabilités potentiellement induites par la réduction. La méthode est robuste dans le sens où elle permet de traiter n'importe quelle erreur modèle et n'importe quelle erreur d'observation, pourvu que ces dernières soient bornées. Nous proposons une forme algorithmique qui permet d'appliquer le filtre à des modèles de grande dimension. L'algorithme calcule l'estimateur minimax ainsi que l'ensemble des états admissibles.

Mots-clés : minimax, réduction, équations différentielles algébriques, estimation, filtrage

Contents

1 Introduction **4**

2 Notation **5**

3 Extended minimax state estimation **6**
3.1 Minimax filter for Ordinary Differential Equations with discrete time 6
3.2 Minimax filter for Differential-Algebraic Equations with discrete time 7
3.3 The case of the non-singular gain 8

4 Model reduction **10**
4.1 Classical reduction 10
4.1.1 Instability of the reachability set 11
4.2 Generalized reduction by means of DAE 13
4.2.1 Linear case 13
4.3 Extended minimax state estimation for DAE 16
4.3.1 Notes about the reduction 16

5 Algorithm and computations **17**
5.1 Derivation of the computational form for the gain G_t 17
5.1.1 First form compatible with computations in high dimension 18
5.1.2 Second form compatible with computations in high dimension, for the gain 19
5.2 Algorithm 19

6 Consistency with Kalman filter **21**

A Sherman-Morrison-Woodbury formula **22**

1 Introduction

Numerical modeling of complex systems such as the Earth's atmosphere involves complex numerical models relying on systems of coupled Partial Differential Equations (PDEs). As an example, consider chemistry-transport models that describe the fate of the pollutants in the atmosphere (e.g., the models described in [Mallet et al., 2007]). For these models, the dimension of the state vector[1] can reach 10^7 or even more, and the time integration has such a large computational cost that only the equivalent of a few dozens of model calls may be affordable. The computational costs of these models and their dimensions raise specific issues when one wants to reduce simulation errors (caused by imperfect model formulation or uncertain inputs) through assimilation of the observed data (sparse observations of the model's state) into the model. Classical assimilation algorithms such as the Kalman filters [Balakrishnan, 1984] can be so demanding in terms of computations that they cannot be applied to these models without a reduction.

Reduced Kalman filters have been developed to address this issue by introducing a reduction of in the filtering algorithm—see [Wu et al., 2008] for an application to the aforementioned chemistry-transport models. In these filters, the key reduction lies in the propagation of the state error covariance matrix which is intractable[2] in Kalman filter. A popular reduced Kalman filter is the so-called ensemble Kalman filter in which the state error covariance matrix is approximated by the empirical variance of the ensemble [Heemink et al., 2001]. The particles can be deterministically sampled like in the SEIK versions[Pham, 2001], in the unscented Kalman filter [Julier and Uhlmann, 1997] or in its reduced version [Moireau and Chapelle, 2010]. Another example is the reduced-rank square-root Kalman filter based on propagation of the most important modes [Verlaan and Heemink, 1995] of the error covariance matrix.

Another direction is the reduction of the model itself and subsequent application of an appropriate filtering technique to the resulting low-dimensional model. The Galerkin projection represents one of the most used techniques for model reduction [Brenner and Scott, 2005]. The idea is to find a low dimensional subspace in the model state space and to restrict the model onto that subspace. Of course, there is a loss of information due to restricting the dynamics of the model onto the subspace. One way to minimize the loss is to generate the subspace by means of the Proper Orthogonal Decomposition (POD) [Homescu et al., 2005].

In this report, we introduce a reduced minimax filter, designed to be applied to high-dimensional models. Minimax approach allows 1) to filter out any model error with bounded energy and observational error either deterministic with bounded energy or stochastic with bounded variance., 2) to estimate the worst-case error and 3) to assess how accurate the link between the model and observed phenomena is. Our approach is to make a reduction of the model itself and to apply the minimax filtering to the reduced model, provided uncertain model error and observation noise are elements of a given bounding set. We introduce a reduced state equation projecting the full state vector onto a subspace which can be generated, e.g., by means of POD. The projection introduces errors

[1] State vector of the PDE after discretization in space.

[2] Since the propagation involves twice as much calls to the tangent linear model as components in the state

that can lead to a reduced state equation with unstable dynamics. In order to address this issue, we introduce an additional energy constraint on the reduced state in the form of a linear algebraic equation. Finally, our reduced model is represented by a Differential-Algebraic Equation (DAE), composed of a reduced state equation and of a linear constraint. We apply an extended version of the minimax filter for DAE [Zhuk, 2010] to the reduced model without further reduction on the filter.

The report is organized as follows. After the notation is introduced in section 2, the extended minimax filter, without reduction, is presented in section 3. This section quickly explains the minimax framework, introduces the filter and comments on the intractability of the computations. The reduction procedure is then derived in section 4. The classical Galerkin projection is first commented. The DAE approach is then introduced, in the linear case and in an extended version for the non-linear case. This section also comments the generation of the projection operator. In section 5, we derive a computational form for the DAE minimax filter presented in section 3. The purpose of the derivations is to provide a form compatible with computations in high dimension. A parallel with Kalman filter can be found in section 6 where it is shown that the minimax filter coincides with the Kalman filter under given assumptions (in particular, without reduction).

2 Notation

Let $\mathcal{M}_t : \mathbb{R}^N \to \mathbb{R}^N$ define the model at some time step $t \in \{0, \dots, T-1\}$:

$$x_{t+1} = \mathcal{M}_t(x_t) + e_t\,, \quad x_0 = x_0^g + e\,, \tag{1}$$

where x_0^g is an approximation of the initial condition with error $e \in \mathbb{R}^N$, $x_t \in \mathbb{R}^N$ denotes the state vector, $e_t \in \mathbb{R}^N$ is the model error.

Let $y_t \in \mathbb{R}^m$ denote the observation of the true state x_t at time t. We assume that y_t satisfies

$$y_t = \mathcal{H}_t(x_t) + \eta_t\,, \tag{2}$$

where $\mathcal{H}_t : \mathbb{R}^N \to \mathbb{R}^m$ is the observation operator mapping the state space into observation space, and $\eta_t \in \mathbb{R}^m$ is the observation error.

We assume that the error (e, e_t, η_t) is uncertain but bounded so that

$$\langle Q^{-1}(e-\overline{e}), e-\overline{e}\rangle + \sum_{t=0}^{T-1} \langle Q_t^{-1}(e_t-\overline{e}_t), e_t-\overline{e}_t\rangle + \sum_{t=0}^{T} \langle R_t^{-1}(\eta_t-\overline{\eta}_t), \eta_t-\overline{\eta}_t\rangle \leqslant 1\,, \tag{3}$$

where $Q, Q_t \in \mathbb{R}^{N\times N}$ and $R \in \mathbb{R}^{m\times m}$ are symmetric positive-definite matrices, and $\overline{e}, \overline{e}_t \in \mathbb{R}^N$ and $\overline{\eta}_t \in \mathbb{R}^m$ may be viewed as systematic errors.

The tangent linear model is $M_t = \mathrm{D}\mathcal{M}_t(x_t) \in \mathbb{R}^{N\times N}$. Consistently we introduce the associated tangent linear operator $H_t = \mathrm{D}\mathcal{H}_t(x_t) \in \mathbb{R}^{m\times N}$.

The reduction applies to the model state, and the reduced model state is denoted $z_t = F_t^T x_t \in \mathbb{R}^n$, with $n \ll N$. $F_t \in \mathbb{R}^{N\times n}$ is a reduction matrix. The minimax estimator of x_t is denoted $\widehat{x}_t \in \mathbb{R}^N$. The minimax estimator is derived from the reduced minimax estimator with $\widehat{x}_t = F_t \widehat{z}_t$.

The tangent linear operators along the trajectory $\widehat{x}_t$ are denoted $\widehat{M}_t = \mathrm{D}\mathcal{M}_t(\widehat{x}_t)F_t \in \mathbb{R}^{N\times n}$ (for $t \geqslant 0$) and $\widehat{H}_t = \mathrm{D}\mathcal{H}_t(\mathcal{M}_{t-1}(\widehat{x}_{t-1}) + \overline{e}_{t-1})F_t \in \mathbb{R}^{m\times n}$

(for $t > 0$), for the model and the observation operator respectively. We also define $\widehat{H}_0 = \mathrm{D}\mathcal{H}_0(x_0^g + \overline{e})F_0 \in \mathbb{R}^{m\times n}$.

We denote by A^+ the Moore-Penrose pseudoinverse of a matrix A. $A^{\frac{1}{2}}$ is the square root of A and satisfies $A = A^{\frac{1}{2}}A^{\frac{T}{2}}$, where $A^{\frac{T}{2}}$ is the transpose of $A^{\frac{1}{2}}$. $I_{k\times k}$ denotes the identity matrix in $\mathbb{R}^{k\times k}$. $\langle\cdot,\cdot\rangle$ denotes the canonical inner product of the Euclidean space.

3 Extended minimax state estimation

3.1 Minimax filter for Ordinary Differential Equations with discrete time

In what follows we present a minimax state estimation algorithm that solves the following filtering problem: given a sequence of observed data $y_0, \ldots, y_T$ in the form (2) and given the uncertainty description (3), one should estimate the state x_T of (1).

Our approach is based on the following idea: to describe how the model propagates uncertain parameters verifying (3). The key point is to construct the so-called reachability set $\mathcal{R}_t$ at time t, that is, the set of all states x_t satisfying (1) and compatible with the description of uncertain parameters (3) and the observed data y_t in the form (2). In other words, the state x_t^* belongs to $\mathcal{R}_t$ if and only if there is a sequence $E^* := (e^*, \{e_0^*, \ldots, e_{t-1}^*\}, \{\eta_0^*, \ldots, \eta_t^*\})$ verifying (3) such that the sequence $x_0^*, \ldots, x_t^*$ computed from (1) for $e = e^*$ and $e_s = e_s^*$, $0 \leqslant s < t$, is compatible with observed data $y_0, \ldots, y_t$ through (2) with $\eta_s = \eta_s^*$, $0 \leqslant s \leqslant t$. This suggests a way to estimate how the model propagates uncertain parameters (initial-condition error e and model error e_t): it is sufficient to have a description of the dynamics of $\mathcal{R}_t$ in time. The true state can only lie in $\mathcal{R}_t$. Note that the dynamics of $\mathcal{R}_t$ takes into account only those realizations of e, e_t which are compatible with the actual realization of observed data $y_0, \ldots, y_t$. Consequently, if $\mathcal{R}_t$ is empty, one can conclude that the errors were wrongly described by (3).

Any point of $\mathcal{R}_t$ can be the true state. In order to obtain a minimax estimate of this true state, we assign to a point $x \in \mathcal{R}_t$ a worst-case error, that is the maximal distance between x and other points of $\mathcal{R}_t$. The point with minimal worst-case error[3] is the minimax estimate $\widehat{x}_t$. Roughly speaking, the worst-case error can be thought of as a "the longest axis" of the minimal ellipsoid containing $\mathcal{R}_t$, and the minimax estimate $\widehat{x}_t$ is the central point of that ellipsoid.

The basics of the minimax state estimation were developed by [Bertsekas and Rhodes, 1971], [Milanese and Tempo, 1985], [Chernousko, 1994], [Kurzhanski and Vályi, 1997], [Nakonechny, 2004]. The main advantages of minimax estimates are as follows: (1) the possibility to filter out any model error and observation noise with bounded energy, (2) the estimation of the worst-case error, (3) fast estimation algorithms in the form of filters, (4) the possibility to evaluate the model, that is to assess how good the model describes observed phenomena.

In this subsection, we assume that $F_t = I_{N\times N}$. In this case, there is no reduction and the state estimation algorithm operates on the full model. Following [Zhuk, 2010], we introduce an extended version of the linear minimax

[3] This point will coincide with the Tchebysheff center of the smallest ellipsoid containing the reachability set.

state estimate $\hat{x}_t$, a minimax gain G_t and the reachability set $\mathcal{R}_t$ for (1):

$$
\begin{aligned}
&G_0 = Q_0^{-1} + \hat{H}_0^T R_0^{-1} \hat{H}_0, \\
&\hat{x}_0 = G_0^{-1} \left(Q_0^{-1} \overline{e} + \hat{H}_0^T R_0^{-1} (y_0 - \overline{\eta}_0) \right), \\
&\beta_0 = \langle R_0^{-1} (y_0 - \overline{\eta}_0), y_0 - \overline{\eta}_0 \rangle, \\
&G_{t+1} = Q_t^{-1} - Q_t^{-1} \widehat{M}_t B_t \widehat{M}_t^T Q_t^{-1} + \hat{H}_{t+1}^T R_{t+1}^{-1} \hat{H}_{t+1}, \\
&B_t = (G_t + \widehat{M}_t^T Q_t^{-1} \widehat{M}_t)^{-1}, \\
&\hat{x}_{t+1}^f = \mathcal{M}_t(\hat{x}_t), \\
&\hat{x}_{t+1} = \hat{x}_{t+1}^f + G_{t+1}^{-1} \hat{H}_{t+1}^T R_{t+1}^{-1} [y_{t+1} - \overline{\eta}_{t+1} - \hat{H}_{t+1} \hat{x}_{t+1}^f] + G_{t+1}^{-1} (Q_t + \widehat{M}_t G_t^{-1} \widehat{M}_t^T)^{-1} \overline{e}_t, \\
&\mathcal{X}_{t+1} = \{ w : \langle G_{t+1} w, w \rangle \leqslant 1 \}, \\
&\beta_{t+1} = \beta_t + \langle R_{t+1}^{-1} (y_{t+1} - \overline{\eta}_{t+1}), y_{t+1} - \overline{\eta}_{t+1} \rangle - \langle B_t^+ G_t^{-1} \hat{x}_t, G_t^{-1} \hat{x}_t \rangle + \langle \widehat{M}_t^T Q_t \widehat{M}_t \overline{e}_t, \overline{e}_t \rangle, \\
&\mathcal{R}_t = \hat{x}_t + \sqrt{1 - \beta_t + \langle G_t \hat{x}_t, \hat{x}_t \rangle} \mathcal{X}_t \,.
\end{aligned}
\tag{4}
$$

Here $\mathcal{R}_t$ denotes an ellipsoidal approximation of the reachability set for the model (1) and β_t is a scaling factor. The dynamics of $\mathcal{X}_t$ describes how the model $\mathcal{M}_t$ propagates uncertain parameters from the bounding set (3) compatible with observed data y_t. The observation-dependent scaling factor β_t defines whether $\mathcal{X}_t$ shrinks or expands. If $1 - \beta_t + \langle G_t \hat{x}_t, \hat{x}_t \rangle < 0$, then the observed data is incompatible with our assumption on uncertainty description (3). In the form (4), the minimax state estimation algorithm can be applied to non-linear models, hence we refer to it as an extended minimax filter. Nevertheless, the theory only supports the algorithm in the linear case—that is, the reachability set is known to contain all possible true states only in the linear case.

The algorithm is far too expensive for high-dimensional systems: it requires to propagate a minimax gain $G_t \in \mathbb{R}^{N \times N}$, where N is the dimension of the state space of the model (1). For instance, with $N = 10^7$ like in air quality applications, the dimension of G_t is $10^7 \times 10^7$, which cannot be manipulated by modern computers because of huge computational loads and out-of-reach memory requirements. Hence a reduction is necessary to carry out the computations for high dimensional systems.

3.2 Minimax filter for Differential-Algebraic Equations with discrete time

A more general form of the filter was derived in [Zhuk, 2010] for DAE problems. The filter addresses the problem

$$
F_{t+1} z_{t+1} = \mathcal{M}_t(F_t z_t) + r_t \,, \quad F_0 z_0 = F_0 F_0^T (x_0^g + e) \,, \quad y_t = \mathcal{H}_t(F_t z_t) + \eta_t \,, \tag{5}
$$

with

$$
\begin{aligned}
\langle Q^{-1}(e - \overline{e}), e - \overline{e} \rangle + \sum_{t=0}^{T-1} \langle Q_t^{-1} (r_t - \overline{r}_t), r_t - \overline{r}_t \rangle \\
+ \sum_{t=0}^{T} \langle R_t^{-1} (\eta_t - \overline{\eta}_t), \eta_t - \overline{\eta}_t \rangle \leqslant 1 \,.
\end{aligned}
\tag{6}
$$

Here $F_t \in \mathbb{R}^{N\times n}$ can be any rectangular matrix and $z_t \in \mathbb{R}^n$ denotes the state of the DAE. If $F_t = I_{N\times N}$, the problem statement is the same as in section 3.1.

Following [Zhuk, 2010], we consider the equation for the minimax gain G_t, for any time $t \in \{0,\dots,T-1\}$:

$$\begin{aligned} B_t &= \left(G_t + \widehat{M}_t^T Q_t^{-1}\widehat{M}_t\right)^+ , \\ G_{t+1} &= F_{t+1}^T\left[Q_t^{-1} - Q_t^{-1}\widehat{M}_t B_t \widehat{M}_t^T Q_t^{-1}\right]F_{t+1} + \hat{H}_{t+1}^T R_{t+1}^{-1}\hat{H}_{t+1}\,, \end{aligned} \tag{7}$$

with the following initialization:

$$G_0 = F_0^T Q^{-1} F_0 + \hat{H}_0^T R_0^{-1}\hat{H}_0\,. \tag{8}$$

For any time $t \in \{0,\dots,T\}$, the minimax estimator is defined as

$$\hat{z}_t = G_t^+ v_t\,, \tag{9}$$

with

$$v_0 = F_0^T Q^{-1}\overline{e} + \hat{H}_0^T R_0^{-1}(y_0 - \overline{\eta}_0)\,, \tag{10}$$

and, for $t \in \{1,\dots,T\}$,

$$\begin{aligned} v_t &= F_t^T Q_{t-1}^{-1}\mathcal{M}_{t-1}(F_{t-1}B_{t-1}v_{t-1}) \\ &+ F_t^T\left[Q_{t-1}^{-1} - Q_{t-1}^{-1}\widehat{M}_{t-1}B_{t-1}\widehat{M}_{t-1}^T Q_{t-1}^{-1}\right]\overline{r}_{t-1} \\ &+ \widehat{H}_t^T R_t^{-1}(y_t - \overline{\eta}_t). \end{aligned} \tag{11}$$

For any time $t \in \{0,\dots,T\}$, the reachability set $\mathcal{R}_t$ is defined as

$$\mathcal{R}_t = \hat{z}_t + \sqrt{1-\beta_t + \langle G_t\hat{z}_t,\hat{z}_t\rangle}\mathcal{X}_t, \quad \mathcal{X}_t = \{x : \langle G_t x, x\rangle \leqslant 1\} \tag{12}$$

with β_t being a scaling factor depending on observations:

$$\beta_{t+1} = \beta_t + \langle R_{t+1}^{-1}(y_{t+1}-\overline{\eta}_{t+1}), y_{t+1}-\overline{\eta}_{t+1}\rangle - \langle B_t^+ G_t^{-1}\hat{z}_t, G_t^{-1}\hat{z}_t\rangle + \langle \widehat{M}_t^T Q_t \widehat{M}_t\overline{r}_t,\overline{r}_t\rangle$$

We have that the reachability set is a translation of the set $\mathcal{X}_t$ induced by the minimax gain G_t. The shape of $\mathcal{X}_t$ depends only on the model, observation operator and bounding set. $\mathcal{X}_t$ describes how the model propagates uncertain initial condition and model error from the bounding set (6). In contrast to the case of ODE, G_t could be singular so that $\mathcal{X}_t$ contains the kernel of G_t. In fact, the part of the system state lying in that kernel is not observable.

3.3 The case of the non-singular gain

Assume for simplicity that $\overline{r}_t = 0$ and $\overline{e} = 0$. Let us further assume that G_t is positive definite for all time istants t. This is the case when, for instance, F_t, for $t \in \{0,T\}$, is of full column rank, or $F_t^T F_t + \hat{H}_t^T\hat{H}_t$ is positive definite. If G_t is positive definite, then $Q_t + \widehat{M}_t G_t^{-1}\widehat{M}_t^T$ is positive definite, and according to Sherman-Morrison-Woodbury formula (see section A), its inverse can be written in the form

$$\left(Q_t + \widehat{M}_t G_t^{-1}\widehat{M}_t^T\right)^{-1} = Q_t^{-1} - Q_t^{-1}\widehat{M}_t(G_t + \widehat{M}_t^T Q_t^{-1}\widehat{M}_t)^{-1}\widehat{M}_t^T Q_t^{-1}\,. \tag{13}$$

Using this identity and the gain equation (7), it is possible to write G_{t+1} as

$$G_{t+1} = F_{t+1}^T \left(Q_t + \widehat{M}_t G_t^{-1} \widehat{M}_t^T\right)^{-1} F_{t+1} + \hat{H}_{t+1}^T R_{t+1}^{-1} \hat{H}_{t+1}\,, \tag{14}$$

which gives an alternative form to the filter. It also proves that G_t is positive definite for all $t \in \{0, \ldots, T\}$.

The state estimator can be rewritten so that the model is applied directly to $F_t \hat{z}_t$ instead of $F_t B_t v_t$. Although it is an equivalent formulation in the linear case, it can make a huge difference when the model is non-linear. In addition, this alternative form makes it easier to interpret the action of the filter. Starting from $\hat{z}_{t+1} = G_{t+1}^{-1} v_{t+1}$ (equation (9)) and the expression (11) for v_{t+1}:

$$\hat{z}_{t+1} = G_{t+1}^{-1} F_{t+1}^T Q_t^{-1} \mathcal{M}_t(F_t B_t G_t \hat{z}_t) + G_{t+1}^{-1} \hat{H}_{t+1}^T R_{t+1}^{-1} (y_{t+1} - \overline{\eta}_{t+1})\,. \tag{15}$$

In case of a linear model and considering the equation for B_t (7), one gets

$$\begin{aligned} \hat{z}_{t+1} =& G_{t+1}^{-1} F_{t+1}^T Q_t^{-1} \widehat{M}_t \left(G_t + \widehat{M}_t^T Q_t^{-1} \widehat{M}_t\right)^{-1} G_t \hat{z}_t \\ &+ G_{t+1}^{-1} \hat{H}_{t+1}^T R_{t+1}^{-1} (y_{t+1} - \overline{\eta}_{t+1})\,, \end{aligned} \tag{16}$$

which, according to the Sherman-Morrison-Woodbury formula, is equivalent to

$$\begin{aligned} \hat{z}_{t+1} =& G_{t+1}^{-1} F_{t+1}^T Q_t^{-1} \widehat{M}_t \left[G_t^{-1} - G_t^{-1} \widehat{M}_t^T (Q_t + \widehat{M}_t G_t^{-1} \widehat{M}_t^T)^{-1} \widehat{M}_t G_t^{-1}\right] G_t \hat{z}_t \\ &+ G_{t+1}^{-1} \hat{H}_{t+1}^T R_{t+1}^{-1} (y_{t+1} - \overline{\eta}_{t+1})\,, \end{aligned} \tag{17}$$

which gives

$$\begin{aligned} \hat{z}_{t+1} =& G_{t+1}^{-1} F_{t+1}^T Q_t^{-1} \left[I_{N\times N} - \widehat{M}_t G_t^{-1} \widehat{M}_t^T (Q_t + \widehat{M}_t G_t^{-1} \widehat{M}_t^T)^{-1}\right] \widehat{M}_t \hat{z}_t \\ &+ G_{t+1}^{-1} \hat{H}_{t+1}^T R_{t+1}^{-1} (y_{t+1} - \overline{\eta}_{t+1})\,, \end{aligned} \tag{18}$$

and, using $I_{N\times N} = (Q_t + \widehat{M}_t G_t^{-1} \widehat{M}_t^T)(Q_t + \widehat{M}_t G_t^{-1} \widehat{M}_t^T)^{-1}$,

$$\hat{z}_{t+1} = G_{t+1}^{-1} F_{t+1}^T \left(Q_t + \widehat{M}_t G_t^{-1} \widehat{M}_t^T\right)^{-1} \widehat{M}_t \hat{z}_t + G_{t+1}^{-1} \hat{H}_{t+1}^T R_{t+1}^{-1} (y_{t+1} - \overline{\eta}_{t+1})\,.$$

Noting that $x = F_t F_t^T x + (I - F_t F_t^T)x$ and using (14) we write

$$\begin{aligned} \hat{z}_{t+1} &= G_{t+1}^{-1}(G_{t+1} - \hat{H}_{t+1}^T R_{t+1}^{-1} \hat{H}_{t+1}) F_{t+1}^T \mathcal{M}_t(F_t \hat{z}_t) \\ &+ G_{t+1}^{-1} F_{t+1}^T \left(Q_t + \widehat{M}_t G_t^{-1} \widehat{M}_t^T\right)^{-1} (I - F_{t+1} F_{t+1}^T) \mathcal{M}_t(F_t \hat{z}_t) \\ &+ G_{t+1}^{-1} \hat{H}_{t+1}^T R_{t+1}^{-1} (y_{t+1} - \overline{\eta}_{t+1}) \\ &= F_{t+1}^T \mathcal{M}_t(F_t \hat{z}_t) + G_{t+1}^{-1} \hat{H}_{t+1}^T R_{t+1}^{-1} (y_{t+1} - \overline{\eta}_{t+1} - \hat{H}_{t+1} F_{t+1}^T \mathcal{M}_t(F_t \hat{z}_t)) \\ &+ G_{t+1}^{-1} F_{t+1}^T \left(Q_t + \widehat{M}_t G_t^{-1} \widehat{M}_t^T\right)^{-1} (I - F_{t+1} F_{t+1}^T) \mathcal{M}_t(F_t \hat{z}_t) \end{aligned}$$

Finally

$$\begin{aligned} \hat{z}_{t+1} &= F_{t+1}^T \mathcal{M}_t(F_t \hat{z}_t) + G_{t+1}^{-1} \hat{H}_{t+1}^T R_{t+1}^{-1} (y_{t+1} - \overline{\eta}_{t+1} - \hat{H}_{t+1} F_{t+1}^T \mathcal{M}_t(F_t \hat{z}_t)) \\ &+ G_{t+1}^{-1} F_{t+1}^T \left(Q_t + \widehat{M}_t G_t^{-1} \widehat{M}_t^T\right)^{-1} (I - F_{t+1} F_{t+1}^T) \mathcal{M}_t(F_t \hat{z}_t) \end{aligned} \tag{19}$$

4 Model reduction

We introduce a reduction method which generalizes the classical Galerkin approach. In the later approach, the model state is projected onto a lower-dimensional subspace so that the dynamics of the full state is represented with a small number of scalars. However this reduction can loose some properties of the full model. For instance, the reduced state equation can introduce instabilities that are not in the full model.

Assume that for each time step t, we have a matrix $F_t \in \mathbb{R}^{n\times n}$ whose columns are linearly-independent orthonormal vectors—we therefore have $F_t^T F_t = I_{n\times n}$. We denote by $\mathcal{F}_t$ the linear span of the columns of F_t. The reduction consists in projecting the true state x_t onto this subspace $\mathcal{F}_t$. We introduce $z_t = F_t^T x_t$, which is the vector of the coefficients of the projection of x_t. Consequently we approximate x_t with $F_t z_t$.

4.1 Classical reduction

The main idea of the classical reduction based on the Galerkin projection is to derive the equation for z_t multiplying (1) by F_{t+1}^T:

$$z_{t+1} = F_{t+1}^T x_{t+1} = F_{t+1}^T \mathcal{M}_t(x_t) + F_{t+1}^T e_t \,. \tag{20}$$

Recalling the definition of z_t we obtain

$$z_{t+1} = F_{t+1}^T \mathcal{M}_t(F_t z_t) + F_{t+1}^T e_t + F_{t+1}^T \mathcal{M}_t(x_t) - F_{t+1}^T \mathcal{M}_t(F_t z_t) \,. \tag{21}$$

Let us define

$$p_t = e_t + \mathcal{M}_t(x_t) - \mathcal{M}_t(F_t F_t^T x_t) \,, \tag{22}$$

so that

$$z_{t+1} = F_{t+1}^T \mathcal{M}_t(F_t z_t) + F_{t+1}^T p_t \,, \quad z_0 = F_0^T (x_0^g + e) \,. \tag{23}$$

p_t is the sum of the model error and a reduction error. If we were to apply the extended minimax filter on the reduced state equation (23), we would need to evaluate the range of values that p_t can take. Since p_t is state dependent and since the true state is unknown, it is hard to determine the range of p_t. The natural approach to suppress the state dependence is to bound the reduction error for all plausible states. Hence we may assume that

$$\|p_t\| \leqslant \|e_t\| + \delta_t \,,$$

where, for instance, δ_t is guaranteed to exist for Lipschitz continuous models provided $F_t F_t^T x_t$ approximates x_t with finite error[4]. With possibly modified Q, Q_t and R_t, we write

$$\langle Q^{-1}(e-\overline{e}), e-\overline{e}\rangle + \sum_{t=0}^{T-1} \langle Q_t^{-1}(p_t-\overline{p}_t), (p_t-\overline{p}_t)\rangle + \sum_{t=0}^{T} \langle R_t^{-1}(\eta_t-\overline{\eta}_t), \eta_t-\overline{\eta}_t\rangle \leqslant 1$$

for p_t defined by (22) and some $\overline{p}_t$ defined as a systematic error of the new model error. Note that only $F_{t+1}^T p_t$ has an impact onto dynamics of z_t. Noting that

$$F_{t+1}^T p_t = F_{t+1}^T F_{t+1} F_{t+1}^T p_t \,,$$

[4] If $\|x_t - \tilde{x}_t\| \leqslant \varepsilon$ and κ is the model Lipschitz constant, we can take $\delta_t = \kappa\varepsilon$.

we see that it is enough to have a bound on $F_{t+1}F_{t+1}^T p_t$ only. Thus we can consider an ellipsoid in the form

$$\begin{aligned}\langle Q^{-1}(e-\overline{e}), e-\overline{e}\rangle + \sum_{t=0}^{T-1}\langle (F_t^T Q_t^{-1} F_t)F_t^T(p_t-\overline{p}_t), F_t^T(p_t-\overline{p}_t)\rangle \\ + \sum_{t=0}^{T}\langle R_t^{-1}(\eta_t-\overline{\eta}_t), \eta_t-\overline{\eta}_t\rangle \leqslant 1\,.\end{aligned} \tag{24}$$

Now we stress that the above procedure could lead to the overestimation of the reachability set of the reduced model (23). This is a consequence of the model reduction (we replace $\mathcal{M}_t$ with $F_{t+1}^T\mathcal{M}_t(F_t)$) and the suppression of state-dependence in the reduction error.

Now let us consider an example illustrating the instabilities that can occur because of the reduction.

4.1.1 Instability of the reachability set

Take a linear model $M = \left[\begin{smallmatrix} 1 & 1 \\ -\frac{1}{2} & 0 \end{smallmatrix}\right]$ and $x_0^g = 0$. Then the state equation (1) for $x_t = (x_t^{(1)}, x_t^{(2)})^T$ reads

$$\begin{aligned} x_{t+1}^{(1)} &= x_t^{(1)} + x_t^{(2)} + e_t^{(1)}, \quad x_0^{(1)} = e^{(1)}, \\ x_{t+1}^{(2)} &= -\frac{1}{2}x_t^{(1)} + e_t^{(2)}, \quad x_0^{(2)} = e^{(2)}, \end{aligned} \tag{25}$$

Assume $Q = Q_t = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$ and $\overline{e} = \overline{e}_t = 0$. For simplicity assume that we have no observations, that is, $\mathcal{H}_t(x) = 0$ and $y_t = 0$. Since all admissible noises η_t must verify $0 = y_t - \mathcal{H}_t(x_t) = \eta_t$ we can deduce that the reachability set is defined in this case by the ellipsoid

$$(e^{(1)})^2 + (e^{(2)})^2 + \sum_{t=0}^{T-1}(e_t^{(1)})^2 + \sum_{t=0}^{T-1}(e_t^{(2)})^2 \leqslant 1 \tag{26}$$

Let us compute the reachability set for (25) at time step $T = 2$. To do so we note that

$$\begin{aligned} x_2^{(1)} &= \frac{e^{(1)}}{2} + e^{(2)} + e_0^{(1)} + e_0^{(2)} + e_1^{(1)}, \\ x_2^{(2)} &= -\frac{e_0^{(1)} + e^{(1)} + e^{(2)}}{2} + e_1^{(2)} \end{aligned}$$

Define $\check{e}_{T-1} = (e, e_0, \ldots, e_{T-1})^T$, so that $\check{e}_1 = (e^{(1)}, e^{(2)}, e_0^{(1)}, e_0^{(2)}, e_1^{(1)}, e_1^{(2)})^T$ and set $\ell_1 = (\frac{1}{2}, 1, 1, 1, 1, 0)^T$, $\ell_2 = (-\frac{1}{2}, -\frac{1}{2}, -\frac{1}{2}, 0, 0, 1)^T$. Then

$$x_2^{(1)} = \langle \ell_1, \check{e}_1\rangle, \quad x_2^{(2)} = \langle \ell_2, \check{e}_1\rangle\,.$$

We note that $\|\check{e}_1\|_2^2 \leqslant 1$ if the components of $\check{e}_1$ verify the inequality (26). Now we can write

$$\max_{\check{e}_1} |x_2^{(i)}| = \max_{\|\check{e}_1\|_2 \leqslant 1} \langle \ell_i, \check{e}_1\rangle = \|\ell_i\|_2,\ i = 1, 2\,,$$

so that

$$\max_{\breve{e}_1} |x_2^{(1)}| = \sqrt{\frac{17}{4}}, \quad \max_{\breve{e}_1} |x_2^{(2)}| = \sqrt{\frac{7}{4}}. \tag{27}$$

By analogy we find

$$\max_{\breve{e}_0} |x_1^{(2)}| = \sqrt{\frac{5}{4}}, \tag{28}$$

Take $F_t = \left[\begin{smallmatrix} 1 \\ 0 \end{smallmatrix}\right]$. The reduced model (23) reads

$$z_{t+1} = z_t + F^T p_t, \quad z_0 = e^{(1)}, \tag{29}$$

with $p_t = (x_t^{(2)} + e_t^{(1)}, e_t^{(2)})^T$. Hence

$$z_{t+1} = z_t + p_t^{(1)}, \tag{30}$$

with $p_t^{(1)} = x_t^{(2)} + e_t^{(1)}$. Note that the differential connection between $x_t^{(2)}$ and $e_t^{(1)}$ (represented by the second equation of (25)) is lost in the reduced model. In other words, we consider $x_t^{(2)}$ from now as a part of the model error. As it was already mentioned above, $(0, e_t^{(2)})^T = (I - F_t F_t^T)p_t$ has no impact on the dynamics of z_t. Let us find a bound on $p_t^{(1)}$. By Minkowski inequality,

$$\sum_{t=0}^{T-1} (x_t^{(2)} + e_t^{(1)})^2 \leqslant q_2(T-1) := \left(1 + \sqrt{\max_{\breve{e}_{T-2}} \sum_{t=0}^{T-1} (x_t^{(2)})^2}\right)^2,$$

where the maximum over $\breve{e}_{T-1}$ is taken so as to define a state-independent ellipsoid. The bounding set for the reduced model (29) may be written as:

$$(e^{(1)})^2 + \sum_{t=0}^{T-1} (p_t^{(1)})^2 \leqslant 1 + q_2(T-1) \tag{31}$$

Let us take $T > 0$ and define $p_t^{(1)} = \sqrt{\frac{1+q_2(T-1)}{T}}$. Then $(e^{(1)}, p_0^{(1)}, \dots p_{T-1}^{(1)})^T$ is admissible, provided $e^{(1)} = 0$. We find from (29) that $z_T = \sqrt{T(1 + q_2(T-1))}$. It is clear that, for $T = 2$,

$$q_2(T-1) = \left(1 + \sqrt{\max_{\breve{e}_0} \sum_{t=0}^{1} (x_t^{(2)})^2}\right)^2 \geqslant \left(1 + \sqrt{\max_{\breve{e}_0} |x_1^{(2)}|}\right)^2.$$

Using (27)–(28) we find

$$z_2 = \sqrt{T(1 + q_2(T-1))} \geqslant \sqrt{2(1 + \left(1 + \sqrt{\frac{5}{4}}\right)^2)} \approx 3.312 > \max |x_2^{(1)}| = \sqrt{\frac{17}{4}} \approx 2.062$$

After two time steps, we overestimated the projection of the true reachability set by more than 60%. Even in this simple case, the reduction error, whose norm must be bound by a state-independent value, can lead to a large reachability set. Also, while the full model is stable, the reduced model is not and the errors will accumulate in time (see (30)).

4.2 Generalized reduction by means of DAE

Above we have seen that the state estimation problem for (23) could be affected by instability of the reduced model so that the reachability set could rapidly expand in time although the reachability set of the full model behaves differently. In what follows we propose a way to further constraint the size of the reachability set for the reduced state, while relying on the same reduction-error estimations as previously.

Consider the reduced model

$$z_{t+1} = F_{t+1}^T \mathcal{M}_t(F_t z_t) + F_{t+1}^T p_t, \quad z_0 = F_0^T(x_0^g + e)\,, \tag{32}$$

and the associated error description

$$\begin{aligned} &\langle Q^{-1}(e-\overline{e}), e-\overline{e}\rangle + \sum_{t=0}^{T-1} \langle (F_t^T Q_t^{-1} F_t) F_t^T (p_t - \overline{p}_t), F_t^T(p_t - \overline{p}_t)\rangle \\ &+ \sum_{t=0}^{T} \langle R_t^{-1}(\eta_t - \overline{\eta}_t), \eta_t - \overline{\eta}_t\rangle \leqslant 1\,. \end{aligned} \tag{33}$$

We introduce an additional constraint onto the reduced state:

$$\sum_{t=0}^{T-1} \langle S_t^{-1} L_t z_t, L_t z_t\rangle \leqslant 1 \tag{34}$$

where S_t^{-1} is a $s \times s$-symmetric positive-definite matrix defining the shape of the bounding set for the reduced state, and $L_t \in \mathbb{R}^{s\times n}$ is a design parameter allowing to constraint a desired part of the reduced state or just a linear combination of the reduced state's components. We do not impose any conditions on L_t. Now we note that the energy constraint can be incorporated into (32)–(33) using the following construction:

$$z_{t+1} = F_{t+1}^T \mathcal{M}_t(F_t z_t) + F_{t+1}^T p_t, \quad z_0 = F_0^T(x_0^g + e), \quad L_t z_t + w_t = 0\,, \tag{35}$$

with

$$\begin{aligned} \langle Q^{-1}(e-\overline{e}), e-\overline{e}\rangle + &\sum_{t=0}^{T-1} \langle (F_t^T Q_t^{-1} F_t) F_t^T (p_t - \overline{p}_t), F_t^T(p_t - \overline{p}_t)\rangle \\ &+ \sum_{t=0}^{T} \langle R_t^{-1}(\eta_t - \overline{\eta}_t), \eta_t - \overline{\eta}_t\rangle \\ &+ \sum_{t=0}^{T-1} \langle S_t^{-1}(w_t - \overline{w}_t), (w_t - \overline{w}_t)\rangle \leqslant 1 + 1 \end{aligned} \tag{36}$$

where $\overline{w}_t$ is a parameter.

4.2.1 Linear case

Let us consider one way to choose L_t in the linear case: $\mathcal{M}_t = M_t$. Consider the DAE in the form

$$F_{t+1} z_{t+1} = M_t F_t z_t + r_t\,, \quad F_0 z_0 = F_0 F_0^T(x_0^g + e)\,. \tag{37}$$

We recall that $F_{t+1}^T F_{t+1} = I_{n\times n}$ and that the solution of a linear algebraic equation $F_{t+1}z = b$ is given by $F_{t+1}^T b$, provided $(I - F_{t+1}F_{t+1}^T)b = 0$. According to this, we write:

$$\begin{aligned} z_{t+1} &= F_{t+1}^T M_t F_t z_t + F_{t+1}^T r_t\,, \quad z_0 = F_0^T(x_0^g + e),\\ 0 &= (I - F_{t+1}F_{t+1}^T)M_t F_t z_t + (I - F_{t+1}F_{t+1}^T)r_t\,. \end{aligned} \tag{38}$$

In order to explore the connection between (32) and (38), we define

$$r_t = F_{t+1}F_{t+1}^T(e_t + M_t x_t) - M_t F_t F_t^T x_t\,. \tag{39}$$

As $F_{t+1}^T F_{t+1} F_{t+1}^T = F_{t+1}^T$ we see that $F_{t+1}^T r_t$ is equal to $F_{t+1}^T p_t$ defined by (22). Since the structure of the first equation of (32) coincides with the structure of the first equation in (38), we conclude that the solutions also coincide, provided p_t is defined by (22) and r_t is defined by (39). Note that for r_t defined by (39), the algebraic constraint in (38) is also satisfied.

r_t can be written in the form

$$r_t = F_{t+1}F_{t+1}^T p_t + (I - F_{t+1}F_{t+1}^T)M_t F_t F_t^T x_t\,. \tag{40}$$

The first term is the sum of the projected model error and the reduction error (see the discussion about p_t in section 4.1). The second term represents the information about the part of the model lying in the orthogonal completion of the reduction subspace. Equations (38) reduce to (32) if $L_t := (I - F_{t+1}F_{t+1}^T)M_t F_t$ and $w_t = L_t F_t^T x_t$. We choose S_t^{-1} so that

$$\sum_{t=0}^{T-1} \langle S_t^{-1} L_t F_t^T x_t, L_t F_t^T x_t\rangle \leqslant 1\,,$$

which allows to keep information about the dynamics of $(I - F_{t+1}F_{t+1}^T)x_{t+1}$. Indeed,

$$(I - F_{t+1}F_{t+1}^T)x_{t+1} = (I - F_{t+1}F_{t+1}^T)M_t(F_t F_t^T x_t + (I - F_t F_t^T)x_t) + (I - F_{t+1}F_{t+1}^T)e_t\,,$$

so that

$$L_t F_t^T x_t = (I - F_{t+1}F_{t+1}^T)x_{t+1} - \left[(I - F_{t+1}F_{t+1}^T)M_t(I - F_t F_t^T)x_t + (I - F_{t+1}F_{t+1}^T)e_t\right]\,.$$

Note that $(I - F_{t+1}F_{t+1}^T)x_{t+1}$ corresponds to the part of the state which is suppressed by the reduction as: $x_{t+1} = F_{t+1}z_{t+1} + (I - F_{t+1}F_{t+1}^T)x_{t+1}$.

The bounding set is defined with (36), which introduces a link between the reduced state and the full state through $(I - F_{t+1}F_{t+1}^T)x_{t+1}$. This allows to limit the artificial increase of the reachability set due to the reduction.

Let us illustrate this feature on the previous example. We have $M = \left[\begin{smallmatrix} 1 & 1 \\ -\frac{1}{2} & 0 \end{smallmatrix}\right]$ and $x_0^g = 0$. Then (38) reads

$$\begin{aligned} z_{t+1} &= z_t + r_t^{(1)}, \quad z_0 = e^{(1)},\\ 0 &= -\frac{1}{2}z_t + r_t^{(2)}\,. \end{aligned} \tag{41}$$

Define $L_t = -\frac{1}{2}$. Now we need to define S_i^{-1} from the condition (34). To do so we define

$$q_1(T) := \left(1 + \sqrt{\max_{\check{e}_{T-1}} \sum_{t=0}^{T} (x_t^{(1)})^2}\right)^2\,.$$

Let us compute $q_1(T)$ and $q_2(T)$ for a given $T > 0$. Define $X(T-1) := (x_0^{(1)}, x_0^{(2)}, x_1^{(1)}, x_1^{(2)}, \ldots, x_{T-1}^{(1)}, x_{T-1}^{(2)})^T$, $X_1(T-1) = (x_0^{(1)}, x_1^{(1)}, \ldots, x_{T-1}^{(1)})^T$, $X_2(T-1) = (x_0^{(2)}, x_1^{(2)}, \ldots, x_{T-1}^{(2)})^T$ and

$$A = \begin{pmatrix} I_{2\times 2} & 0 & \ldots & 0 \\ -M & I_{2\times 2} & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & \ldots & -M & I_{2\times 2} \end{pmatrix}$$

Also let us introduce a linear mapping π_i defined by the rule $\pi_i(X(T-1)) = X_i(T-1)$, $i = 1, 2$. Then $X(T-1) = A\check{e}_{T-2}$ as it follows from (25) so that $X_i(T-1) = \pi_i(A\check{e}_{T-2})$. We have

$$\sum_{t=0}^{T-1} (x_t^{(i)})^2 = \langle A^T \pi_i^T \pi_i A \check{e}_{T-2}, \check{e}_{T-2} \rangle, \quad i = 1, 2\,,$$

so that

$$\max_{\check{e}_{T-2}} \sum_{t=0}^{T-1} (x_t^{(i)})^2 = \max_{\check{e}_{T-2}} \langle A^T \pi_i^T \pi_i A \check{e}_{T-2}, \check{e}_{T-2} \rangle, \quad i = 1, 2\,,$$

and since $\|\check{e}_{T-2}\|_2^2 \leqslant 1$ due to (26) we find that

$$\max_{\check{e}_{T-2}} \langle A^T \pi_i^T \pi_i A \check{e}_{T-2}, \check{e}_{T-2} \rangle = \lambda_i^{\max}(A^T \pi_i^T \pi_i A), \quad i = 1, 2\,,$$

where $\lambda_i^{\max}(A^T \pi_i^T \pi_i A)$ denotes the maximal eigenvalue of the matrix $A^T \pi_i^T \pi_i A$, $i = 1, 2$. Recalling that

$$q_i(T-1) := \left(1 + \sqrt{\max_{\check{e}_{T-2}} \sum_{t=0}^{T-1} (x_t^{(i)})^2}\right)^2, \quad i = 1, 2\,,$$

we find

$$q_i(T-1) = \left(1 + \sqrt{\lambda_i^{\max}(A^T \pi_i^T \pi_1 A)}\right)^2, \quad i = 1, 2\,.$$

Straightforward computation shows that

$$\lambda_1^{\max}(A^T \pi_1^T \pi_1 A) \approx 8.58, \quad \lambda_2^{\max}(A^T \pi_2^T \pi_2 A) \approx 2.83$$

for $T = 4$. Using the above formulae we find that

$$q_1(3) \approx 15.44, \quad q_2(3) \approx 7.19$$

Setting $S_j = \frac{q_1(3)}{4}$ we obtain:

$$\sum_{j=0}^{3} S_j^{-1} L_j^2 z_j^2 \leqslant 1$$

Now, taking into account the bounding set (31), we obtain that (36) has the following form

$$(e^{(1)})^2 + \sum_{j=0}^{3} (r_j^{(1)})^2 + \sum_{j=0}^{3} S_j^{-1} L_j^2 z_j^2 \leqslant 1 + q_2(3) + 1$$

We claim that $r_t^{(1)} = \sqrt{\frac{q_2(3)+1}{4}}$, $e^{(1)} = 0$ is not admissible for $T = 4$ in contrast to the bounding set (26) considered above. To see this we note that $\sum_{j=0}^{3}(r_j^{(1)})^2 = q_2(3)+1$ so that it is enough to show that $\sum_{j=0}^{3} S_j^{-1}L_j^2 z_j^2 > 1$. We find from (41) that

$$z_j^2 = j^2(r_0^{(1)})^2 = j^2\frac{q_2(3)+1}{4}\,,$$

so that

$$\sum_{j=0}^{3} S_j^{-1}L_j^2 z_j^2 = \sum_{j=0}^{3} j^2\frac{(q_2(3)+1)}{4q_1(3)} \approx 1.858 > 1\,.$$

Finally we see that the solution z_t of the reduced state equation (41) corresponding to $r_t^{(1)} = \sqrt{\frac{q_2(3)+1}{4}}$ and $e^{(1)} = 0$ is not admissible starting from $T = 4$. In contrast, it is admissible for any T in the case of the reduced model (29).

4.3 Extended minimax state estimation for DAE

After the considerations of the previous section, we introduced the following filtering problem:

$$\begin{aligned}
&z_{t+1} = F_{t+1}^T\mathcal{M}_t(F_t z_t) + F_{t+1}^T p_t,\\
&z_0 = F_0^T(x_0^g + e), \quad L_t z_t + w_t = 0,\\
&\langle Q^{-1}(e-\overline{e}), e-\overline{e}\rangle + \sum_{t=0}^{T}\langle R_t^{-1}(\eta_t-\overline{\eta}_t), \eta_t-\overline{\eta}_t\rangle \qquad (42)\\
&+ \sum_{t=0}^{T-1}\langle (F_t^T Q_t^{-1}F_t)F_t^T p_t, F_t^T p_t\rangle + \sum_{t=0}^{T-1}\langle S_t^{-1}w_t, w_t\rangle \leqslant 1+1\,.
\end{aligned}$$

We define a descriptor matrix $\widetilde{F} = \left[\begin{smallmatrix} I_{n\times n} \\ 0 \end{smallmatrix}\right]$. We can extend the model and its associated error matrix: the new DAE model is $\widetilde{\mathcal{M}_t} = \left[\begin{smallmatrix} F_t^T\mathcal{M}_t F_t \\ L_t \end{smallmatrix}\right]$ and $\widetilde{Q}_t = \left[\begin{smallmatrix} F_t^T Q_t^{-1}F_t & 0 \\ 0 & S_t^{-1} \end{smallmatrix}\right]$. With these definition, we can apply the (extended) DAE minimax filter from section 3.2, simply by substituting F_t with $\widetilde{F}$, $\mathcal{M}_t$ with $\widetilde{\mathcal{M}_t}$ and Q_t with $\widetilde{Q}_t$. Also R_t should be modified to take into account the additional error due to reduction in the observation equation, since $\mathcal{H}_t(F_t z_t)$ is involved instead of $\mathcal{H}_t(x_t)$. The computational version of this algorithm is presented in the section 5.

4.3.1 Notes about the reduction

The reduction can be seen as a projection of the full model state onto the subspace $\mathcal{F}_t$ spanned by the columns of F_t. The columns of F_t can be determined in a number of ways, but the usual reduction approach is based on Proper Orthogonal Decomposition (POD), also called principal component analysis. In order to carry out a POD, we propose the following procedure.

The simulation period $[0, T]$ is split into sub-periods $[0, T_1[$, $[T_1, T_2[$, ... Within a sub-period $[T_i, T_{i+1}[$, before any filtering takes place, the model can be run and a sequence of full states $(\widetilde{x}_{T_i}, \ldots, \widetilde{x}_{T_{i+1}})$ is generated[5]. A POD is

[5] These states can be seen as forecasts.

carried out on this sequence, so that the reduction retains the dynamics of the model.

Over the period $[0, T]$, this algorithm produces matrices F_t that only depend on the sub-period: for any $t_1, t_2 \in [T_i, T_{i+1}[$, $F_{t_1} = F_{t_2}$. Let us focus on a sub-period during which the reduction matrix is constant in time: $F_t = F$. Consider the definition (39) of the error in the linear case. It is decomposed in the projection of the model error $FF^T e_t$ and the difference $FF^T M_t x_t - M_t FF^T x_t$ that is the commutation error between FF^T and M_t. If the projector FF^T commutes with the model, then no error due to the reduction can accumulate in time. Otherwise, some of the model dynamics is lost because of the reduction.

Another source of error may lie at the transitions between two sub-periods. There are changes in the matrix F_t that lead to additional errors, even if the projectors $F_t F_t^T$ and $F_{t+1} F_{t+1}^T$ commute with M_t: $r_t = F_{t+1} F_{t+1}^T e_t + (F_{t+1} F_{t+1}^T - F_t F_t^T) M_t x_t$.

The POD of $(\tilde{x}_{T_i}, \ldots, \tilde{x}_{T_{i+1}})$ may be efficient to reproduce the dynamics of the model, but it may be irrelevant for the assimilation of the observations. In (19), the second term on the right-hand side is a correction that involves the discrepancy between observations and model state. If this discrepancy is not in the subspace $\mathcal{F}_t$, the correction can be removed. For instance, if $y_{t+1} - \overline{\eta}_{t+1} - \widehat{H}_{t+1} F_{t+1}^T \mathcal{M}_t(F_t \hat{z}_t) \notin \mathcal{F}_t$ and $R_{t+1} = I_{m \times m}$, then the correction term is zero, which means that the observations have no impact on the state estimate (except indirectly through G_{t+1}). It is therefore advised to take this issue into account, e.g., by using the POD of $(\tilde{x}_{T_i}, \ldots, \tilde{x}_{T_{i+1}}, y_{T_i} - \overline{\eta}_{T_i} - \widehat{H}_{T_i} F_{T_i}^T \tilde{x}_{T_i}, \ldots, y_{T_{i+1}} - \overline{\eta}_{T_{i+1}} - \widehat{H}_{T_{i+1}} F_{T_{i+1}}^T \tilde{x}_{T_{i+1}})$.

5 Algorithm and computations

5.1 Derivation of the computational form for the gain G_t

Either in (7) or in (14), the gain may not be computed directly because of the large computational costs. If the dimension N of the state space is high, e.g., $N = 10^7$, the inversion of Q_t in (7) and the inversion of $Q_t + \widehat{M}_t G_t^{-1} \widehat{M}_t^T$ in (14) cannot be performed due to memory requirements and computational cost. In order to make the computations tractable, the filter's matrices should be transformed.

A key step is the representation of the matrices Q_t. In practice, it is expected that they are approximated in square root form: $Q_t \simeq Q_t^{\frac{1}{2}} Q_t^{\frac{T}{2}}$, where $Q_t^{\frac{1}{2}}$ is composed of $q \ll N$ columns, and $Q_t^{\frac{T}{2}} = \left(Q_t^{\frac{1}{2}}\right)^T$. As a consequence, some directions in the state space are assumed to be perfectly known (no uncertainty), which is usually not a realistic assumption. In addition, $Q_t^{\frac{1}{2}} Q_t^{\frac{T}{2}}$ is singular. In order to circumvent this issue, it is reasonable to introduce

$$Q_t = Q_t^{\frac{1}{2}} Q_t^{\frac{T}{2}} + D_t \tag{43}$$

where D_t is a positive definite diagonal matrix. The diagonal elements of D_t are likely to be small. They essentially acknowledge the fact that every component in the space state is associated with at least a small error.

One needs to make sure that for any errors satisfying the inequality (3) with the exact Q_t, the inequality still holds after Q_t is replaced with its approximation. At least, the ellipsoid induced by the approximation should contain the ellipsoid induced by the exact model error description. It is always possible to make the ellipsoid bigger with a larger D_t, so that it is always possible to rely on the approximation (43) without underestimating the model error.

5.1.1 First form compatible with computations in high dimension

In the form (14), the main computational issue is the inversion of

$$Q_t + \widehat{M}_t G_t^{-1} \widehat{M}_t^T = Q_t^{\frac{1}{2}} Q_t^{\frac{T}{2}} + D_t + \widehat{M}_t G_t^{-1} \widehat{M}_t^T \,. \tag{44}$$

Introducing the positive-definite matrix $J_t = D_t + \widehat{M}_t G_t^{-1} \widehat{M}_t^T$ and using the Sherman-Morrison-Woodbury formula (see section A), the inversion of (44) can be written as

$$\left(Q_t + \widehat{M}_t G_t^{-1} \widehat{M}_t^T\right)^{-1} = J_t^{-1} - J_t^{-1} Q_t^{\frac{1}{2}} \left(I_{q\times q} + Q_t^{\frac{T}{2}} J_t^{-1} Q_t^{\frac{1}{2}}\right)^{-1} Q_t^{\frac{T}{2}} J_t^{-1} \,. \tag{45}$$

Applying the Sherman-Morrison-Woodbury formula to J_t^{-1} gives

$$\left(D_t + \widehat{M}_t G_t^{-1} \widehat{M}_t^T\right)^{-1} = D_t^{-1} - D_t^{-1} \widehat{M}_t \left(G_t + \widehat{M}_t^T D_t^{-1} \widehat{M}_t\right)^{-1} \widehat{M}_t^T D_t^{-1} \,. \tag{46}$$

We introduce the following scaled matrices

$$\begin{aligned} \breve{F}_{t+1} &= D_t^{-\frac{1}{2}} F_{t+1} \in \mathbb{R}^{N\times n} \,, \\ \breve{Q}_t^{\frac{1}{2}} &= D_t^{-\frac{1}{2}} Q_t^{\frac{1}{2}} \in \mathbb{R}^{N\times q} \,, \\ \breve{U}_t &= D_t^{-\frac{1}{2}} \widehat{M}_t \left(G_t + \widehat{M}_t^T D_t^{-1} \widehat{M}_t\right)^{-\frac{1}{2}} \in \mathbb{R}^{N\times n} \,. \end{aligned} \tag{47}$$

Then $J_t^{-1} = D_t^{-\frac{1}{2}} (I_{N\times N} - \breve{U}_t \breve{U}_t^T) D_t^{-\frac{1}{2}}$. With this expression of J_t^{-1} and the identity (45), the expression (14) of G_t reads

$$\begin{aligned} G_{t+1} = {} & \breve{F}_{t+1}^T (I_{N\times N} - \breve{U}_t \breve{U}_t^T) \breve{F}_{t+1} \\ & - \breve{F}_{t+1}^T (I_{N\times N} - \breve{U}_t \breve{U}_t^T) \breve{Q}_t^{\frac{1}{2}} \left(I_{q\times q} + Q_t^{\frac{T}{2}} J_t^{-1} Q_t^{\frac{1}{2}}\right)^{-1} \times \\ & \times \breve{Q}_t^{\frac{T}{2}} (I_{N\times N} - \breve{U}_t \breve{U}_t^T) \breve{F}_{t+1} + \hat{H}_{t+1}^T R_{t+1}^{-1} \hat{H}_{t+1} \,, \end{aligned} \tag{48}$$

and

$$\begin{aligned} G_{t+1} = {} & \breve{F}_{t+1}^T \breve{F}_{t+1} - \breve{F}_{t+1}^T \breve{U}_t \left(\breve{F}_{t+1}^T \breve{U}_t\right)^T \\ & - \left[\breve{F}_{t+1}^T \breve{Q}_t^{\frac{1}{2}} - \breve{F}_{t+1}^T \breve{U}_t \breve{U}_t^T \breve{Q}_t^{\frac{1}{2}}\right] \left(I_{q\times q} + Q_t^{\frac{T}{2}} J_t^{-1} Q_t^{\frac{1}{2}}\right)^{-1} \times \\ & \times \left[\breve{F}_{t+1}^T \breve{Q}_t^{\frac{1}{2}} - \breve{F}_{t+1}^T \breve{U}_t \breve{U}_t^T \breve{Q}_t^{\frac{1}{2}}\right]^T + \hat{H}_{t+1}^T R_{t+1}^{-1} \hat{H}_{t+1} \,. \end{aligned} \tag{49}$$

Let $V_t = I_{q\times q} + Q_t^{\frac{T}{2}} J_t^{-1} Q_t^{\frac{1}{2}}$, then

$$V_t = I_{q\times q} + \breve{Q}_t^{\frac{T}{2}} \breve{Q}_t^{\frac{1}{2}} - \breve{Q}_t^{\frac{T}{2}} \breve{U}_t \breve{U}_t^T \breve{Q}_t^{\frac{1}{2}} \in \mathbb{R}^{q\times q} \tag{50}$$

is symmetric positive-definite (because J_t^{-1} is symmetric positive-definite). Finally,

$$\begin{aligned} G_{t+1} =& \check{F}_{t+1}^T \check{F}_{t+1} - \check{F}_{t+1}^T \check{U}_t \left(\check{F}_{t+1}^T \check{U}_t\right)^T \\ & - \left[\check{F}_{t+1}^T \check{Q}_t^{\frac{1}{2}} - (\check{F}_{t+1}^T \check{U}_t)(\check{U}_t^T \check{Q}_t^{\frac{1}{2}})\right] V_t^{-\frac{1}{2}} V_t^{-\frac{T}{2}} \left[\check{F}_{t+1}^T \check{Q}_t^{\frac{1}{2}} - (\check{F}_{t+1}^T \check{U}_t)(\check{U}_t^T \check{Q}_t^{\frac{1}{2}})\right]^T \\ & + \hat{H}_{t+1}^T R_{t+1}^{-1} \hat{H}_{t+1} \,. \end{aligned} \tag{51}$$

The matrix G_0 can be computed as

$$G_0 = \check{F}_0^T \check{F}_0 - \check{F}_0^T \check{Q}^{\frac{1}{2}} \left[I_{q\times q} + \check{Q}^{\frac{T}{2}} \check{Q}^{\frac{1}{2}}\right]^{-1} \check{Q}^{\frac{T}{2}} \check{F}_0 + \hat{H}_0^T R_0^{-1} \hat{H}_0 \,. \tag{52}$$

In the forms (52) and (51), the computation of G_t, $t \in \{0, \ldots, T\}$, is tractable. In (51), the rounded brackets indicate in what order the matrix multiplications should be carried out. With these indications, the largest multiplications require nqN or n^2N operations. Note that the operation $\widehat{M}_t^T D_t^{-1}$ or $D_t^{-1}\widehat{M}_t$ is simply the multiplication of every row of $\widehat{M}_t$ with a diagonal element of D_t^{-1}. Three matrix inversions and three square root decompositions are needed:

1. $D_t^{-\frac{1}{2}}$ which is trivial since D_t is diagonal;
2. $\left(G_t + \widehat{M}_t^T D_t^{-1} \widehat{M}_t\right)^{-\frac{1}{2}}$ which is tractable because the matrix is of small size—it is in $\mathbb{R}^{n\times n}$;
3. $V_t^{-\frac{1}{2}}$ which also involves a small matrix, in $\mathbb{R}^{q\times q}$.

In (52), the matrix $I_{q\times q} + \check{Q}^{\frac{T}{2}} \check{Q}^{\frac{1}{2}}$ is of small size as well.

The algorithm corresponding to the filter in this form is shown in section 5.2.

5.1.2 Second form compatible with computations in high dimension, for the gain

A derivation similar to that of section 5.1.1 is possible, but starting from (7) directly. In this approach, Q_t^{-1} is decomposed with the Sherman-Morrison-Woodbury formula:

$$Q_t^{-1} = D_t^{-1} - D_t^{-1} Q_t^{\frac{1}{2}} \left(I_{q\times q} + Q_t^{\frac{T}{2}} D_t^{-1} Q_t^{\frac{1}{2}}\right)^{-1} Q_t^{\frac{T}{2}} D_t^{-1} \,. \tag{53}$$

Injecting this expression for Q_t^{-1} in (7) leads to another tractable form of reduced minimax filter.

5.2 Algorithm

The algorithm presented below is valid for the case when the gain G_t is non-singular. Also, for simplicity, we assume that $\overline{r}_t = 0$. Based on the filter in the form of section 3.3, the algorithm reads:

Parameters x_0^g, $\overline{e}$, $Q = Q^{\frac{1}{2}} Q^{\frac{T}{2}} + D$, $\mathcal{M}_t$, F_t, $Q_t = Q_t^{\frac{1}{2}} Q_t^{\frac{T}{2}} + D_t$, y_t, $\mathcal{H}_t$, $\overline{\eta}_t$, R_t.

Initialization

$$
\begin{aligned}
\breve{F}_0 &= D^{-\frac{1}{2}} F_0 \,, \\
\breve{Q}^{\frac{1}{2}} &= D^{-\frac{1}{2}} Q^{\frac{1}{2}} \,, \\
\hat{H}_0 &= \mathrm{D}\mathcal{H}_0(x_0^g + \overline{e}) F_0 \,, \\
G_0 &= \breve{F}_0^T \breve{F}_0 - \left(\breve{F}_0^T \breve{Q}^{\frac{1}{2}}\right) \left[I_{q\times q} + \breve{Q}^{\frac{T}{2}} \breve{Q}^{\frac{1}{2}}\right]^{-1} \left(\breve{F}_0^T \breve{Q}^{\frac{1}{2}}\right)^T + \hat{H}_0^T R_0^{-1} \hat{H}_0 \,, \\
v_0 &= \breve{F}_0^T D^{-\frac{1}{2}} (x_0^g + \overline{e}) - \left(\breve{F}_0^T \breve{Q}^{\frac{1}{2}}\right) \left[I_{q\times q} + \breve{Q}^{\frac{T}{2}} \breve{Q}^{\frac{1}{2}}\right]^{-1} \breve{Q}^{\frac{T}{2}} D^{-\frac{1}{2}} (x_0^g + \overline{e}) + \hat{H}_0^T R_0^{-1} (y_0 - \overline{\eta}_0) \,, \\
\hat{z}_0 &= G_0^{-1} v_0 \,.
\end{aligned}
$$

For $t \in \{0, \ldots, T-1\}$

$$
\begin{aligned}
\widetilde{M}_t &= D_t^{-\frac{1}{2}} \mathrm{D}\mathcal{M}_t(\hat{z}_t) \,, \\
\breve{U}_t &= \widetilde{M}_t \left(G_t + \widetilde{M}_t^T \widetilde{M}_t\right)^{-\frac{1}{2}} \,, \\
\breve{F}_{t+1} &= D_t^{-\frac{1}{2}} F_{t+1} \,, \\
\breve{Q}_t^{\frac{1}{2}} &= D_t^{-\frac{1}{2}} Q_t^{\frac{1}{2}} \,, \\
\hat{H}_{t+1} &= \mathrm{D}\mathcal{H}_{t+1}(\mathcal{M}_t(F_t \hat{z}_t) + \overline{r}_t) F_{t+1} \,, \\
V_t &= I_{q\times q} + \breve{Q}_t^{\frac{T}{2}} \breve{Q}_t^{\frac{1}{2}} - \left(\breve{Q}_t^{\frac{T}{2}} \breve{U}_t\right) \left(\breve{Q}_t^{\frac{T}{2}} \breve{U}_t\right)^T \,, \\
G_{t+1} &= \breve{F}_{t+1}^T \breve{F}_{t+1} - \left(\breve{F}_{t+1}^T \breve{U}_t\right) \left(\breve{F}_{t+1}^T \breve{U}_t\right)^T \\
&\quad - \left[\breve{F}_{t+1}^T \breve{Q}_t^{\frac{1}{2}} - (\breve{F}_{t+1}^T \breve{U}_t)(\breve{U}_t^T \breve{Q}_t^{\frac{1}{2}})\right] V_t^{-1} \left[\breve{F}_{t+1}^T \breve{Q}_t^{\frac{1}{2}} - (\breve{F}_{t+1}^T \breve{U}_t)(\breve{U}_t^T \breve{Q}_t^{\frac{1}{2}})\right]^T \\
&\quad + \hat{H}_{t+1}^T R_{t+1}^{-1} \hat{H}_{t+1} \,, \\
\hat{z}_{t+1} &= F_{t+1}^T \mathcal{M}_t(F_t \hat{z}_t) + G_{t+1}^{-1} \hat{H}_{t+1}^T R_{t+1}^{-1} (y_{t+1} - \overline{\eta}_{t+1} - \hat{H}_{t+1} F_t^T \mathcal{M}_t(F_t \hat{z}_t)) \\
&\quad + G_{t+1}^{-1} \breve{F}_{t+1}^T D_t^{\frac{1}{2}} (I - F_{t+1} F_{t+1}^T) \mathcal{M}_t(F_t \hat{z}_t) \\
&\quad - G_{t+1}^{-1} (\breve{F}_{t+1}^T \breve{U}_t) \breve{U}_t^T D_t^{\frac{1}{2}} (I - F_{t+1} F_{t+1}^T) \mathcal{M}_t(F_t \hat{z}_t) \\
&\quad - G_{t+1}^{-1} \left[\breve{F}_{t+1}^T \breve{Q}_t^{\frac{1}{2}} - (\breve{F}_{t+1}^T \breve{U}_t)(\breve{U}_t^T \breve{Q}_t^{\frac{1}{2}})\right] V_t^{-\frac{1}{2}} \times \\
&\quad \times V_t^{-\frac{T}{2}} \left[\breve{Q}_t^{\frac{T}{2}} - (\breve{Q}_t^{\frac{T}{2}} \breve{U}_t) \breve{U}_t^T\right] D_t^{\frac{1}{2}} (I - F_{t+1} F_{t+1}^T) \mathcal{M}_t(F_t \hat{z}_t) \,.
\end{aligned}
$$

In order to derive the equation for $\hat{z}_t$, we transformed $F_{t+1}^T \left(Q_t + \widehat{M}_t G_t^{-1} \widehat{M}_t^T\right)^{-1} \widehat{M}_t$ found in equation (19) in the same way as $F_{t+1}^T \left(Q_t + \widehat{M}_t G_t^{-1} \widehat{M}_t^T\right)^{-1} \hat{F}_{t+1}$ in the gain in (51).

A similar form of the algorithm is available in the data assimilation library Verdandi (http://verdandi.gforge.inria.fr/).

6 Consistency with Kalman filter

In the linear case ($\mathcal{M}_t(x) = \widehat{M}_t x$), without systematic model error or observational error ($\overline{r}_t = 0$ and $\overline{\eta}_t = 0$) and without reduction ($F_t = I_{n\times n}$), the filter coincides with the Kalman filter. The bound on the error (e, r_t, η_t) (see equation (3)) should be reinterpreted in terms of variances. We introduce a scaling coefficient $\alpha > 0$. It is assumed that the variance of e is αQ, the variance of r_t is αQ_t, and the variance of η_t is αR_t.

At time $t = 0$, from (10), (8) and (9), the minimax estimator is

$$\widehat{x}_0 = \left(Q^{-1} + \widehat{H}_0^T R_0^{-1} \widehat{H}_0\right)^{-1} (Q^{-1}\overline{e} + \widehat{H}_0^T R_0^{-1} y_0)\,, \tag{54}$$

which leads to

$$\begin{aligned}\widehat{x}_0 = &\left(Q^{-1} + \widehat{H}_0^T R_0^{-1} \widehat{H}_0\right)^{-1} \left(Q^{-1} + \widehat{H}_0^T R_0^{-1} \widehat{H}_0 - \widehat{H}_0^T R_0^{-1} \widehat{H}_0\right) \overline{e} \\ &+ \left(Q^{-1} + \widehat{H}_0^T R_0^{-1} \widehat{H}_0\right)^{-1} \widehat{H}_0^T R_0^{-1} y_0\,,\end{aligned}$$

hence

$$\widehat{x}_0 = \overline{e} + \left(Q^{-1} + \widehat{H}_0^T R_0^{-1} \widehat{H}_0\right)^{-1} \widehat{H}_0^T R_0^{-1} (y_0 - \widehat{H}_0 \overline{e})\,, \tag{55}$$

which coincides with the best linear unbiased estimator (BLUE) of x_0, based on the background state $\overline{e}$ with error variance αQ and the observation vector y_0 with error variance αR_0. The gain matrix is $K_0 = (Q^{-1} + \widehat{H}_0^T R_0^{-1} \widehat{H}_0)^{-1} \widehat{H}_0^T R_0^{-1} = Q\widehat{H}_0^T (\widehat{H}_0 Q \widehat{H}_0^T + R_0)^{-1}$. In addition, the error variance of BLUE is $Q^{-1} + \widehat{H}_0^T R_0^{-1} \widehat{H}_0 = G_0^{-1}$.

Let us assume that at time $t-1$, $\widehat{x}_{t-1}$ coincides with the Kalman estimator and that its error variance is G_{t-1}^{-1}. G_t is invertible and the minimax estimator at time t is

$$\widehat{x}_t = G_t^{-1} Q_{t-1}^{-1} \widehat{M}_{t-1} \left(G_{t-1} + \widehat{M}_{t-1}^T Q_{t-1}^{-1} \widehat{M}_{t-1}\right)^{-1} G_{t-1} \widehat{x}_{t-1} + G_t^{-1} H_t^T R_t^{-1} y_t\,. \tag{56}$$

According to (14),

$$G_t = P_t^{-1} + \widehat{H}_t^T R_t^{-1} \widehat{H}_t\,, \tag{57}$$

with

$$P_t = Q_{t-1} + \widehat{M}_{t-1} G_{t-1}^{-1} \widehat{M}_{t-1}^T\,. \tag{58}$$

Note that, in the Kalman framework, αP_t corresponds to the error variance of the forecast $\widehat{M}_{t-1}\widehat{x}_{t-1}$, and αG_t^{-1} corresponds to the error variance of the Kalman estimator.

We now prove that $\widehat{x}_t = G_t^{-1} P_t^{-1} \widehat{M}_{t-1} \widehat{x}_{t-1} + G_t^{-1} H_t^T R_t^{-1} y_t$ with the following derivation (using the Sherman-Morrison-Woodbury formula, see section A):

$$\begin{aligned}&Q_{t-1}^{-1} \widehat{M}_{t-1} \left(G_{t-1} + \widehat{M}_{t-1}^T Q_{t-1}^{-1} \widehat{M}_{t-1}\right)^{-1} G_{t-1} = \\ &= Q_{t-1}^{-1} \widehat{M}_{t-1} \left(G_{t-1}^{-1} - G_{t-1}^{-1} \widehat{M}_{t-1}^T (Q_{t-1} + \widehat{M}_{t-1} G_{t-1}^{-1} \widehat{M}_{t-1}^T)^{-1} \widehat{M}_{t-1} G_{t-1}^{-1}\right) G_{t-1} \\ &= Q_{t-1}^{-1} \widehat{M}_{t-1} \left(I_{n\times n} - G_{t-1}^{-1} \widehat{M}_{t-1}^T P_t^{-1} \widehat{M}_{t-1}\right) \\ &= Q_{t-1}^{-1} \widehat{M}_{t-1} - Q_{t-1}^{-1} (P_t - Q_{t-1}) P_t^{-1} \widehat{M}_{t-1} \\ &= P_t^{-1} \widehat{M}_{t-1}\,.\end{aligned} \tag{59}$$

Therefore, (56) can be rewritten as

$$\hat{x}_t = G_t^{-1} P_t^{-1} \widehat{M}_{t-1} \hat{x}_{t-1} + G_t^{-1} H_t^T R_t^{-1} y_t \,, \tag{60}$$

or

$$\hat{x}_t = \left(P_t^{-1} + \hat{H}_t^T R_t^{-1} \hat{H}_t \right)^{-1} P_t^{-1} \widehat{M}_{t-1} \hat{x}_{t-1} + \left(P_t^{-1} + \hat{H}_t^T R_t^{-1} \hat{H}_t \right)^{-1} H_t^T R_t^{-1} y_t \,. \tag{61}$$

This equation is in the same form as (54), with P_t in place of Q and the forecast $\widehat{M}_{t-1}\hat{x}_{t-1}$ in place of $\bar{e}$. The same derivation gives

$$\hat{x}_t = \widehat{M}_{t-1} \hat{x}_{t-1} + K_t (y_t - H_t \widehat{M}_{t-1} \hat{x}_{t-1}) \,, \tag{62}$$

where we introduce the Kalman gain $K_t = \left(P_t^{-1} + \hat{H}_t^T R_t^{-1} \hat{H}_t \right)^{-1} H_t^T R_t^{-1}$.

This concludes the proof of the consistency between the minimax filter and the Kalman filter under the assumptions previously mentioned. Nevertheless, the meaning of the matrices Q, Q_{t-1} and R_t differ since they are covariance matrices in the Kalman filter while they define bounded errors in the minimax filter. In many practical applications (e.g., in geophysical modeling), the errors are bounded and modeling the errors with normal distributions is not realistic. As a consequence, a typical assumption is that a random variable has a clipped normal distribution and always lies in an interval centered at its expectation and of width equal to four times its standard deviation (where one finds about 95% of the total probability of a normal distribution). In this case, the scaling parameter α would be $\frac{1}{4}$.

Acknowledgement

This work was carried out during the second author's tenure of an ERCIM "Alain Bensoussan" Fellowship Programme.

A Sherman-Morrison-Woodbury formula

The Sherman-Morrison-Woodbury matrix identity is

$$(S + N_1 W N_2)^{-1} = S^{-1} - S^{-1} N_1 (W^{-1} + N_2 S^{-1} N_1)^{-1} N_2 S^{-1} \tag{63}$$

if S and W are nonsingular matrices.

References

[Balakrishnan, 1984] Balakrishnan, A. (1984). *Kalman Filtering Theory*. Opt. Soft., Inc., N.Y.

[Bertsekas and Rhodes, 1971] Bertsekas, D. and Rhodes, I. B. (1971). Recursive state estimation with a set-membership description of the uncertainty. *IEEE Trans. Automat. Contr.*, AC-16:117–128.

[Brenner and Scott, 2005] Brenner, S. and Scott, R. (2005). *The Mathematical Theory of Finite Element Methods*. Springer, 2nd edition. ISBN 0-3879-5451-1.

[Chernousko, 1994] Chernousko, F. L. (1994). *State Estimation for Dynamic Systems* . Boca Raton, FL: CRC.

[Heemink et al., 2001] Heemink, A. W., Verlaan, M., and Segers, A. J. (2001). Variance reduced ensemble Kalman filtering. *Mon. Wea. Rev.*, 129:1,718–1,728.

[Homescu et al., 2005] Homescu, C., Petzold, L. R., and Serban, R. (2005). Error estimation for reduced-order models of dynamical systems. *SIAM J. Numer. Anal.*, 43(4):1,693–1,714.

[Julier and Uhlmann, 1997] Julier, S. J. and Uhlmann, J. K. (1997). A new extension of the Kalman filter to nonlinear systems. In *Proceedings of AeroSense: The 11th International Symposium on Aerospace/Defense Sensing, Simulation and Controls*.

[Kurzhanski and Vályi, 1997] Kurzhanski, A. and Vályi, I. (1997). *Ellipsoidal calculus for estimation and control*. Systems & Control: Foundations & Applications. Birkhäuser Boston Inc., Boston, MA.

[Mallet et al., 2007] Mallet, V., Quélo, D., Sportisse, B., Ahmed de Biasi, M., Debry, É., Korsakissok, I., Wu, L., Roustan, Y., Sartelet, K., Tombette, M., and Foudhil, H. (2007). Technical Note: The air quality modeling system Polyphemus. *Atmos. Chem. Phys.*, 7(20):5,479–5,487.

[Milanese and Tempo, 1985] Milanese, M. and Tempo, R. (1985). Optimal algorithms theory for robust estimation and prediction. *IEEE Trans. Automat. Control*, 30(8):730–738.

[Moireau and Chapelle, 2010] Moireau, P. and Chapelle, D. (2010). Reduced-order unscented Kalman filtering with application to parameter identification in large-dimensional systems. *ESAIM: Control, Opt. and Calc. Var.*

[Nakonechny, 2004] Nakonechny, A. (2004). *Optimal control and estimation for partial differential equations*. National Taras Shevchenko University Publishing. (in Ukrainian).

[Pham, 2001] Pham, D. T. (2001). Stochastic methods for sequential data assimilation in strongly nonlinear systems. *Mon. Wea. Rev.*, 129:1,194–1,207.

[Verlaan and Heemink, 1995] Verlaan, M. and Heemink, A. W. (1995). Reduced rank square root filters for large scale data assimilation problems. In *Second International Symposium on Assimilation of Observations in Meteorology and Oceanography*, pages 247–252, Japan.

[Wu et al., 2008] Wu, L., Mallet, V., Bocquet, M., and Sportisse, B. (2008). A comparison study of data assimilation algorithms for ozone forecasts. *J. Geophys. Res.*, 113(D20310).

[Zhuk, 2010] Zhuk, S. (2010). Minimax state estimation for linear discrete-time differential-algebraic equations. *Automatica J. IFAC*, 46(11):1785–1789.

Centre de recherche INRIA Paris – Rocquencourt
Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex (France)

Centre de recherche INRIA Bordeaux – Sud Ouest : Domaine Universitaire - 351, cours de la Libération - 33405 Talence Cedex
Centre de recherche INRIA Grenoble – Rhône-Alpes : 655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier
Centre de recherche INRIA Lille – Nord Europe : Parc Scientifique de la Haute Borne - 40, avenue Halley - 59650 Villeneuve d'Ascq
Centre de recherche INRIA Nancy – Grand Est : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex
Centre de recherche INRIA Rennes – Bretagne Atlantique : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex
Centre de recherche INRIA Saclay – Île-de-France : Parc Orsay Université - ZAC des Vignes : 4, rue Jacques Monod - 91893 Orsay Cedex
Centre de recherche INRIA Sophia Antipolis – Méditerranée : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399